A network model approach to document retrieval taking into account domain knowledge

Peter Scheir, Stefanie N. Lindstaedt
Know-Center
Inffeldgasse 21a, 8010 Graz, Austria
{pscheir, slind}@know-center.at

Abstract

We preset a network model for context-based retrieval allowing for integrating domain knowledge into document retrieval. Based on the premise that the results provided by a network model employing spreading activation are equivalent to the results of a vector space model, we create a network representation of a document collection for retrieval. We extended this well explored approach by blending it with techniques from knowledge representation. This leaves us with a network model for finding similarities in a document collection by content-based as well as knowledge-based similarities.

1 Introduction

The work presented here originated in the context of the project APOSDLE[1]. One of the objectives of this project is contextualized delivery of information to knowledge workers. For this task we create formal models (task model, domain model, competence model) of the different aspects of the work context to represent the environment the knowledge worker operates in. We use this information, besides content-based analysis of resources, for retrieving information relevant to the current work situation.

When defining a model[2] of the context of a knowledge worker (cf. [Ulbrich *et al.*, 2006]) we noticed that there are three classes[3] of objects that could be used for querying an information retrieval system:

- A *set of concepts* of a knowledge representation that describes the situation of the knowledge worker, for example the current actions a person performs or the competencies he or she acquires.

  This concepts stem from the formal models that are used to represent the context of the knowledge worker.

- A *set of documents* that are related to the current situation of the knowledge worker, for example the document template he or she is currently interacting with, or the process documentation the person is reading.

  In our approach documents are related to concepts from the task and the domain model. This enables us to infer which documents are associated with the current task and vice versa.

- A *set of terms* which are related to his or hers current situation, examples for such terms would be parts of documents the person currently views or a text he or she currently types.

  The set of terms is not related to any of the models that span our context model. Nevertheless we think of it as a vital addition to our approach to retrieval.

To increase the chances for successfully supporting the worker with resources, a model taking all three classes of objects (concepts, terms, documents) as query items into account was needed. In this contribution we present our suggestion for such a model and discuss the technical feasibility of a system implementing the model. Additionally we will present related work in the field of network models in information retrieval.

Our contribution is structured as follows: First (in Section 2) we give an overview on network models in information retrieval, explain underlying concepts such as spreading activation and present systems operating on knowledge representations as well as on document collections. Then in Section 3 we introduce our model and the challenges related to our approach. We will discuss the technical realization of the model, present the lessons we have learned so far and point out the benefits of our approach. We conclude this contribution (Section 4) with a brief discussion and future tasks on our research agenda.

2 Network models in information retrieval

Network models have a long tradition in information retrieval and experienced great popularity in the 1980s, inspired by the rise of neural networks. Systems using network representations often employ a processing technique called spreading activation. Spreading activation originates from cognitive psychology where it servers as mechanism for explaining how knowledge is represented and processed in the human brain (cf. [Anderson, 1983]). The human mind is modelled as a network of nodes, which represent concepts and are connected by edges. Starting from a set of initially activated nodes, activation spreads over the edges to their neighbours. Those nodes with the highest level of energy are seen to be the most similar to the set of nodes activated initially. A detailed introduction to spreading activation in information retrieval can be found in [Crestani, 1997].

As in this section we will give an overview on network models in the field of information retrieval, we find

[1] Advanced Process Oriented and Self-Directed Learning Environment - http://www.aposdle.org/

[2] The context-model used in APOSDLE is based on a meta-model that defines mappings between a task model, a domain model and a competence model. These models are created according to the current application domain of the system.

[3] Currently the deduction of these classes from the meta-model is not described in a formalized way.

it important to note that our view on information retrieval is inspired by Raphael [Raphael, 1968], who sees information retrieval as document retrieval as well as fact retrieval. Therefore we see document-content-based retrieval as well as knowledge-representation-based retrieval as components of information retrieval. In the work presented here we introduce a network model covering both of these aspects for finding resources to support knowledge workers.

Two *divergent* applications of network model in information retrieval exist which we aim to unify in our approach: (1) retrieval in knowledge bases and (2) retrieval in document bases. We will now give a short overview about those applications.

2.1 Knowledge retrieval using network models

There exist several classical and contemporary systems that model a knowledge base as a network of nodes that is searched by spreading activation. Examples are documented in [Cohen and Kjeldsen, 1987] or [Berger *et al.*, 2004]. While knowledge representations are a current issue since the early days of artificial intelligence research, it was the ongoing effort of the Semantic Web community that lead to sound mechanisms for creating knowledge representations in the form of ontologies (see [Guarino, 1998] for an exact definitions of an ontology). New standards for knowledge representation where defined, new methodologies and tools for developing ontologies where developed. Examples of recent systems, employing a network model and spreading activation for identifying similarity in ontologies, exist [Alani *et al.*, 2003] [Rocha *et al.*, 2004].

2.2 Document retrieval using network models

With the rise of neural networks researchers tried applying this paradigm to information retrieval, resulting in a neural network like representation of document collections. One of the first systems employing a network representation for document retrieval was AIR [Belew, 1989]. In AIR the document collection is modelled as a network of nodes. Two classed of nodes exist, one for representing documents and one for representing terms contained in the documents. Some systems for document retrieval also focused on a central aspect of neural network: learning. So incorporate [Belew, 1989] and [Kwok and Grunfeld, 1996] relevance feedback of the user by reweighing connections between documents and terms into the network representation.

3 A network model for integrating domain knowledge into document retrieval

The approach presented here combines two sources for retrieving resources for the knowledge worker. On the one hand content-based similarity of documents is used, while on the other hand similarities stem from a knowledge representation. For integrating the two sources, a network representation is used as this form of model is well suited for both aspects of our approach (cf. Sections 2.1 and 2.2) and allows for the integration of the two aspects. The resulting model can be represented as three layers architecture: (1) A layer for documents, (2) a layer for terms extracted from documents and (3) a layer for concepts originating from the ontology(s) used (see Figure 1).

The document layer and the term layer stem from the document base present in the system. For every document

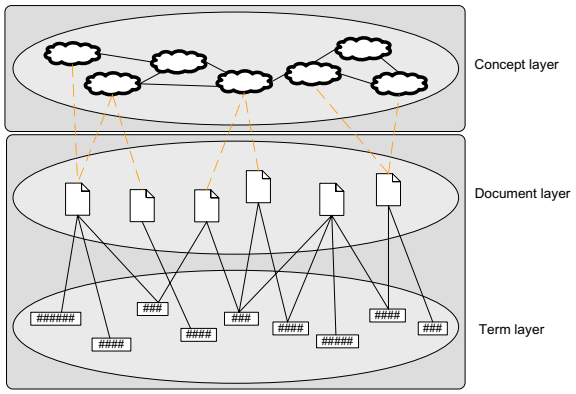

Figure 1: The network model consisting of concept layer, document layer and term layer

present in the document base a document node in the document layer is created. Document nodes feature a unique identifier to be precisely distinguishable. The term nodes result from the terms contained in the documents. Classical text indexing techniques are used to extract the terms from the documents, additional pre-processing steps as stemming can be added to the pre-processing queue, before adding term nodes to the term layer. If a term is contained in a document the term node is connected to the document node by an undirected edge in the network. The edge is weighted by term-frequency - inverse-document-frequency (tf-idf). This data structure can be represented as a matrix similar to the term-document-matrix used in other information retrieval approaches. The document and term layer of the model presented here are comparable to the approaches reviewed in 2.2.

The concept layer is created by transforming a knowledge representation into a network structure. This is done by using the underlying graph structure of the knowledge representation, in which nodes symbolizing concepts are connected by edges. The edges already present in the graph data structure can be weighted using various approaches, one of them would be to take the paths between two nodes into account (cf. [Rocha *et al.*, 2004]). The concept layer is comparable to the systems described in Section 2.1.

Items from every layer can be used to initiate a query. Therefore this layer concept fulfils our requirements of a retrieval system for contextualized information delivery (cf. Section 1), which should be able to build a query from a set of concepts, a set of terms and a set of documents. The network is searched using spreading activation (cf. Section 2). Depending on the current context of a user, a set of concepts nodes, term nodes and documents nodes is activated, when the network is queried. From these nodes energy spreads over the network, leaving those document nodes with the most energy that are closest related to the set of query nodes. Theses nodes will be presented as result of the contextualized retrieval process. As the edges connecting node are undirected activation can spread in both direction over an edge.

3.1 Challenges resulting from this approach

After the creation of the three layers the challenge of combining the already well integrated document and term layer with the concept layer exists (symbolised by the dashed lines in Figure 1). While the document and term layers stem from resources present in the document base of the

system (cf. Section 2.2), the concept layer originates from the knowledge representation(s) used (cf. Section 2.1). Per se those two tiers are not connected, as the domain models are created by experts for a special purpose, while the resources in the document base of the system are created by the workers in the company during their daily job. As manually relating concepts from formal models to documents is a burdensome task an interesting challenge is raised that is similar to one existing when trying to establish a Semantic Web: A lot of resources already exist in the current form of the web, but most of them are not annotated semantically. Therefore our research effort will precisely observe and contribute to current and future developments in fields such as ontology learning [Buitelaar *et al.*, 2005] and the (semi-)automatic semantic annotation of resources with semantic metadata [Handschuh and Staab, 2003].

Additionally, we find it important to integrate support for reasoning over the semantics of edges that stem from a knowledge representation. This is similar to the approach presented in [Wolverton, 1995] where the spread of activation is guided by a reasoning engine that decides which edges to follow depending on the retrieval task.

3.2 Related work

Currently no approaches to information retrieval are known to the authors, which provide information based on content-based similarities of resources and a knowledge representation by combining them into a network model. An effort into this direction was I3R [Croft and Thompson, 1987] where a network structure was introduced allowing for connecting documents, extracted terms and concepts from a semantic network. In the presented prototype only content-based similarity of text documents and similarities based on the same author and co-referenced works were used. In [Agosti and Crestani, 1993] a similar network structure is suggested. It remains unclear in both approaches how content-based similarity and the semantic layer are connected.

3.3 Technical realization

To explore the functionality of our model two core components of our retrieval engine need to be built: (1) An implementation of the spreading activation algorithm is needed and (2) an efficient way of storing the network data-structure used for the retrieval task. We have already implemented two research prototypes for testing purposes (one for the concept layer and one for the term and document layer). While we initially started with an in-memory representation of the whole network structure, formed by objects representing nodes that reference other objects, we have dropped this idea in favour of an search-index-based representation (on hard-disk) of the network. In the following we will explain why:

In our model three types of nodes exist: One for concepts from a knowledge representation, one for document and one for terms. As in our model a document node can be connected to term nodes and concept nodes only, and no connections between terms nodes and terms nodes or documents nodes and documents nodes exist, the graph representation of our model is sparse. For this reason we store our network as an adjacency list as this is the general accepted procedure in managing sparse graphs (cf. [Sedgewick and Schidlowsky, 2003]). This approach equals a large hash table, with source nodes as keys and edge, destination node pairs as values. The index of a standard text search engine like Lucene[4] is well suited to perform the task of storing this adjacency list representation of the network. For example in [Lux and Granitzer, 2005] our colleagues demonstrated how to efficiently use Lucene's index for the task of graph retrieval.

Using the search-index-based approach presented here we are still able to use our existing implementation of spreading activation algorithms, gain the benefit of easy serializability of the network structure and avoid the problem of high memory usage.

3.4 Lessons learned

To assuring the quality of our model at an early stage we did a prototypical implementation of a system, consisting of document nodes and term nodes, employing spreading activation for search. The built system can be compared to the ones described in Section 2.2. Our initial goal was to evaluate our system against the vector space model, similar to the work done by [Salton and Buckley, 1988]. We started building a system with the same indexing back-end as our spreading-activation-based system but with a vector-space-model-based search. During the implementation of the vector space model we were surprised by the similarity of the code for a basic vector space model implementation to the one for a basic the network model using spreading activation. This leads us to further research which we will now briefly summarize:

When [Salton and Buckley, 1988] did the first comparison between the vector space model and network models using spreading activation it ended in favour of the vector space model. We (as [Crestani, 1997]) believe that these results come from the fact that in this comparison a mature version of the vector space model was compared to a rather basic form of the network model. [Wilkinson and Hingston, 1991] present a two layer network model (term and document nodes) which employed the cosine measure from the vector space model by weighting the edges between term and document nodes using tf-idf. This already allowed for speculations on the similarity between the two models. As noted by [Mandl, 2000] in 1994 [Mothe, 1994] demonstrated theoretically and empirically that the results from a network model using spreading activation in the first wave of spreading are identical to those returned by the vector space model, if the same weighting functions are used for both models.

We see this finding as an important quality assurance measure for our work as we can assume that the results of the network model built from documents and terms is equivalent to that of a vector space model if the same weighting techniques are used. We can now extend this model by blending it with a knowledge representation.

3.5 Benefits of our approach

Our approach to context-based retrieval using a network model allows for the integration of domain knowledge in the form of ontologies into document retrieval. In existing systems knowledge representations are statically implemented and are not easily changeable, as it is in our case.

Advantages of the employed approach of search based on a network representation and spreading activation are (cf. [Alves and Jorge, 2005]): The independence of the content type of objects present in the network (but of course different similarities measures have to be defined) and the robustness to missing information. A system built on our

[4] http://lucene.apache.org/

model allows for ostensive [Campbell and van Rijsbergen, 1996] retrieval (a form of retrieval where a user not explicitly formulates a query for a search, but selects material that currently is useful for him), as needed for the user context-based approach.

As spreading activation can be mapped to a distributed web environment using a messaging approach and our approach incorporates knowledge representations with document collections we think that the presented method for retrieval should fit well for searching the Semantic Web.

4 Conclusion and Future Work

We have presented our effort on integrating content-based similarity of documents with a knowledge representation by using a network structure for the task of context-based retrieval. While we are satisfied with our results so far, several issues remain to address: On the theoretical side more research on the integration of the conceptual layer and the document layer (cf. Section 3.1) has to be done, here we want to employ methods from the fields of ontology learning and (semi-)automatic semantic annotation of resources. The more technical part of our future work will focus on the implementation of the retrieval system itself, following the approach presented in the technical realization part of this contribution (Section 3.3).

Acknowledgments

We thank Thomas Mandl and Josiane Mothe for their support on the equivalency of the vector space model and network models using spreading activation. We also thank our colleague Armin Ulbrich for his founding work on the context-model and his helpful comments.

This work has been partially funded under grant 027023 in the IST work programme of the European Community. The Know-Center is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Ministry of Transport, Innovation and Technology (www.ffg.at/index.php?cid=95) and by the State of Styria.

References

[Agosti and Crestani, 1993] Maristella Agosti and Fabio Crestani. A methodology for the automatic construction of a hypertext for information retrieval. In *SAC '93: Proceedings of the 1993 ACM/SIGAPP symposium on Applied computing*. ACM Press, 1993.

[Alani *et al.*, 2003] Harith Alani, Srinandan Dasmahapatra, Kieron O'Hara, and Nigel Shadbolt. Identifying communities of practice through ontology network analysis. *IEEE Intelligent Systems*, 18(2):18–25, 2003.

[Alves and Jorge, 2005] Mario A. Alves and Alipio M. Jorge. Minibrain: a generic model of spreading activation in computers, and example specialisations. In *ECML/PKDD 2005 workshop 'Subsymbolic paradigms for learning in structured domains'*, 2005.

[Anderson, 1983] John R. Anderson. A spreading activation theory of memory. *Journal of Verbal Leaning and Verbal Behaviour*, 22:261–295, 1983.

[Belew, 1989] Richard K. Belew. Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. In *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, 1989.

[Berger *et al.*, 2004] Helmut Berger, Michael Dittenbach, and Dieter Merkl. An adaptive information retrieval system based on associative networks. In *Proceedings of the 1st Asia-Pacific Conference on Conceptual Modelling*, 2004.

[Buitelaar *et al.*, 2005] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini, editors. *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, 2005.

[Campbell and van Rijsbergen, 1996] Iain Campbell and Cornelis J. van Rijsbergen. The ostensive model of developing information needs. In *of 2nd International Conference on Conceptions of Library Science*, 1996.

[Cohen and Kjeldsen, 1987] Paul R. Cohen and Rick Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Inf. Process. Manage.*, 23:255–268, 1987.

[Crestani, 1997] Fabio Crestani. Application of spreading activation techniques in information retrieval. *Artif. Intell. Rev.*, 11:453–482, 1997.

[Croft and Thompson, 1987] W. B. Croft and R. H. Thompson. I3R: a new approach to the design of document retrieval systems. *J. Am. Soc. Inf. Sci.*, 38:389–404, 1987.

[Guarino, 1998] Nicola Guarino. Formal Ontology and Information Systems. In *International Conference On Formal Ontology In Information Systems*, 1998.

[Handschuh and Staab, 2003] Siegfried Handschuh and Steffen Staab, editors. *Annotation for the Semantic Web*. IOS Press, 2003.

[Kwok and Grunfeld, 1996] K.L. Kwok and L. Grunfeld. TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS. In *The Fourth Text REtrieval Conference (TREC-4)*, 1996.

[Lux and Granitzer, 2005] Mathias Lux and Michael Granitzer. A fast and simple path index based retrieval approach for graph based semantic descriptions. In *Proceedings of the Second International Workshop on Text-Based Information Retrieval*, 2005.

[Mandl, 2000] Thomas Mandl. Tolerant and adaptive information retrieval with neural networks. In *Global Dialogue. Science and Technology Thinking the Future at EXPO 2000 Hannover*. 2000.

[Mothe, 1994] Josiane Mothe. Search mechanisms using a neural network model. In *Proceedings of the RIAO 94 (Recherche dInformation assiste par Ordinateur)*, 1994.

[Raphael, 1968] Bertram Raphael. SIR : Semantic Information Retrieval. In *Semantic Information Processing*. MIT Press, Cambridge, MA, 1968.

[Rocha *et al.*, 2004] Cristiano Rocha, Daniel Schwabe, and Marcus Poggi Aragao. A hybrid approach for searching in the semantic web. In *Proceedings of the 13th international conference on World Wide Web*, 2004.

[Salton and Buckley, 1988] G. Salton and C. Buckley. On the use of spreading activation methods in automatic information. In *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, 1988.

[Sedgewick and Schidlowsky, 2003] Robert Sedgewick and Michael Schidlowsky. *Algorithms in Java, Part 5: Graph Algorithms*. Addison-Wesley, 2003.

[Ulbrich *et al.*, 2006] Armin Ulbrich, Peter Scheir, Stefanie N. Lindstaedt, and Manuel Goertz. A context-model for supporting work-integrated learning. In *Innovative Approaches for Learning and Knowledge Sharing - First European Conference on Technology Enhanced Learning*, 2006.

[Wilkinson and Hingston, 1991] Ross Wilkinson and Philip Hingston. Using the cosine measure in a neural network for document retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, 1991.

[Wolverton, 1995] Michael Wolverton. An investigation of marker-passing algorithms for analogue retrieval. In *Case-Based Reasoning Research and Development*, 1995.